DACS TECHNICAL MONOGRAPH SERIES

A COMPARISON OF RADC AND NASA/SEL SOFTWARE DEVELOPMENT DATA

DECEMBER 1982

DACS

Data & Analysis Center for Software

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: TM-1	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): A COMPARISON OF RADC AND NASA/SEL SOFTWARE DEVELOPMENT DATA; TECHNICAL MONOGRAPH SERIES	5. TYPE OF REPORT & PERIOD COVERED: Interim Report Jan. 1981 - May 1981
	6. PERFORMING ORG. REPORT NUMBER: N/A
7. AUTHOR(s): IIT Research Institute	8. CONTRACT OR GRANT NUMBER(s): F30602-78-C-0255
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Data & Analysis Center for Software RADC/ISISI Griffiss AFB, NY 13441	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (COEE) Griffiss AFB, NY 13441	12. REPORT DATE: May 1981, revised Dec. 1982
	13. NUMBER OF PAGES: 27
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer: John Palaimo (COEE)
DACS Source Code No. 413570
Cost: $5.00

Available from:
Data & Analysis Center for Software
RADC/ISISI
Griffiss AFB, NY 13441

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Computer Software
Productivity Factors
Data Analysis
Software Experience Data
Data Repository

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This DACS Technical Monograph examines the effect on seven statistical relationships when the productivity data from the NASA/SEL dataset is combined with the DACS Productivity Dataset. This monograph contains 7 pages of text plus 21 pages of graphs.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

DACS Data & Analysis Center for Software

AN INFORMATION ANALYSIS CENTER

DACS TECHNICAL MONOGRAPH SERIES

A COMPARISON OF RADC AND NASA/SEL SOFTWARE DEVELOPMENT DATA

ORIGINALLY PUBLISHED MAY 1981

REVISED DECEMBER 1982

DACS The Data & Analysis Center for Software

is a Department of Defense Information Analysis Center, operated by IIT Research Institute under contract to the Rome Air Development Center, AFSC.

The Data & Analysis Center for Software (DACS) is a Department of Defense (DoD) information analysis center sponsored by the Air Force Systems Command, Rome Air Development Center (RADC), and operated by IIT Research Institute (IITRI). DACS serves as a central source for current, readily usable data and information concerning software technology.

The major functions of the DACS are: to maintain a computer database of empirical data collected on the development and maintenance of computer software; to produce and distribute subsets of the database for use by software researchers; to maintain a software technology information base of technical documents, project status information, and evaluation data pertinent to the computer software field; to analyze the data and information and produce technical reports; to maintain a current awareness program which includes dissemination of technical information, assessments of technological developments, and publication of a quarterly newsletter; to provide technical assistance in the form of technical information and special studies of topics related to software engineering and software technology.

To obtain more information on the products and services of the DACS, contact:

Data & Analysis Center for Software
RADC/ISISI
Griffiss Air Force Base, NY 13441

Telephone: 315/336-0937
Autovon: 587-3395

PREFACE

This is one of a series of publications dealing with software development and maintenance data. This publication presents the results of an examination of several relationships between the size of a software project and other metrics and rates which describe certain attributes of a software project's development process. Some of these relationships have been previously examined in the Rome Air Development Center (RADC) report by Richard Nelson entitled "Software Data Collection and Analysis." That document is available from the Data & Analysis Center for Software (DACS).

After the establishment of the DACS, the original dataset analyzed by Nelson was augmented by the inclusion of new data. It was then questioned whether it would be better to analyze the data as one pooled dataset or to keep the subsets separated. The present analysis is a first attempt in this direction.

The information presented in this document may prove useful for comparative purposes, as a diverse collection of software projects are examined. This document establishes a procedure by which the DACS may compare future software development data with data presently residing in the DACS software development database. Finally this document begins to examine the benefits and disadvantages of using Modern Programming Practices (MPPs).

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

In September of 1978, Richard Nelson of Rome Air Development Center (RADC) completed a report entitled "Software Data Collection and Analysis" in which he examined several statistical relationships within the RADC Software Productivity Database. The relationships studied attempted to relate the size of a software project with various other metrics describing the development process. The seven primary relationships studied by Nelson are given below:

1) Project Size vs. Productivity (source lines per manmonth)

2) Project Size vs. Development Effort (manmonths)

3) Project Size vs. Development Duration (months)

4) Project Size vs. Average Manloading (manmonths per month)

5) Project Size vs. Number of Errors

6) Project Size vs. Spatial Error Rate (number of errors per 1000 source lines)

7) Project Size vs. Effort Based Error Rate (number of errors per 10 manmonths of development effort)

This report summarizes the results of a similar examination of all but one of these relationships when data from the NASA/SEL database is merged with the RADC data. The relationship between Project Size and Average Manloading will not be examined because of the different methods used in computing this metric for the two databases. However, another possible relationship, given below, is examined.

8) Project Size vs. Temporal Error Rate (number of errors per month of development time)

The data items described in the previously mentioned Nelson report are used again here. A number of the metrics stored in the RADC database were derived from raw data and had to be computed from the data parameters in the NASA/SEL database. These metrics are described in the following paragraphs:

(1) Project Size - This metric is stored in the NASA/SEL database in three ways: total delivered lines of source code; new source code; and modified source code. The total delivered lines of source code is used in this report to obtain the largest possible sample size.

(2) Development Effort - This metric is stored in the NASA/SEL database in terms of hours expended for each of three categories of labor (these being management, programers, and clerical) . These hours were summed over the three categories, and manmonths were computed based on 160 hours in one manmonth.

(3) Development Duration - This metric is computed from the project schedule dates stored in the database and rounded to the nearest whole month.

(4) Number of Errors - This metric is equal to the number of change reports submitted during development which is recorded in the database. This figure corresponds to the metric used in the original report by Richard Nelson, where the number of software problem reports submitted during development was assumed to equal the number of errors which occurred during development.

(5) Error Rates - Each of the error rates is computed directly from the four metrics defined in the above paragraphs.

All of the NASA/SEL projects are primarily in a Higher Order Language (HOL), some using subroutines written in Assembly language. All of the NASA/SEL projects may also be considered as having been developed in an environment where Modern Programming Practices (MPPs) were used.

All of the graphs in this report are identical in format to graphs in the original report by Richard Nelson, and this similarity allows easy comparison between reports. Three of the graphs from that report are reproduced in Appendix A for comparison. All of the references to previous results refer to those results in the Nelson report.

2. RESULTS

Each of the graphs in this paper is a scatter plot, with each point representing a software project. In each graph, data from the RADC productivity database is represented by a plus (+) and data from the NASA/SEL database is represented by a circle (o). The data is plotted on a loglog scale and the axes are annotated accordingly. The least squares equation for the log-linear regression line of best fit is included, along with the standard error for future predictions using the regression equation and the coefficient of correlation for the data. The regression models developed in this report are the same as those in the Nelson report. They differ slightly from Nelson's due to the inclusion of the NASA/SEL data. This regression line and the lines representing one standard error of estimate about the regression are also included on the graph.

Figures 1A, 1B, 1C Productivity vs. Size

Figure 1A represents the relationship of Productivity in lines per manmonth vs. Project Size in delivered source lines for all RADC and NASA/SEL projects regardless of development language or methodology. Figures 1B and 1C represent the same relationship for projects using MPPs and HOL projects using MPPs regardless of language, respectively. Figure A1 in Appendix A, exhibits this relationship for projects where MPPs were not used during development. The sets of data points used in Figure 1B and in Figure A1 in Appendix A are mutually exclusive and these figures may be used for comparison. Notice how the slopes of these two regression equations change significantly and that the exponents acting on the independent variable differ in their signs; note also the low level of correlation, and the relatively large standard error. One interesting observation notable in each of the three graphs in this report, is that the NASA/SEL projects appear more frequently above the average productivity represented by the regression equation, indicating that those projects were generally developed under conditions of higher programmer productivity than projects described in the RADC database. However, the primary conclusion one draws from the regression (and the lack of a pattern to the scatter), is that Productivity correlates very little or not at all with Project Size.

Figures 2A, 2B, 2C Project Effort vs. Size

Figure 2A represents the relationship of Project Effort vs. Project Size in source lines for all RADC and NASA/SEL projects. Figures 2B and 2C represent the same relationship for those projects developed using MPPs and HOL projects developed using MPPs, respectively. Figure A2 in Appendix A represents this relationship for those projects developed without the use of MPPs. The sets of data points used to produce Figures 2B and A2 are mutually exclusive. A comparison of these two regression equations and the standard error associ ed with the first figure shows that, while a change resulted from the change i 'ata sets being examined, these changes in the coefficient and exponents of regression equations, are not very large. Notice that in each graph the majority of NASA projects appear to have taken less effort to deve than the projects of comparable size described in the RADC database. As would be expected, each of the three graphs indicates that Development Effort correlates highly with Project Size. Also, it should be noted that the exponent term of the three regressions does not differ much from 1, indicating that this relationship is nearly linear.

Figures 3A, 3B, 3C Project Duration vs. Size

Figure 3A represents the relationship of Project Duration in months vs. Project Size in source lines for all of the RADC and NASA/SEL projects. As before, projects developed using MPPs, and HOL projects developed using MPPs were examined separately in Figures 3B and 3C respectively. Figure A3 in Appendix A represents this relationship for projects where MPPs were not used during development. Figures 3B and A3 were generated from mutually exclusive data sets, and a comparison of these results indicates that this relationship appears not to be largely affected by the use of MPPs during development. Unlike the previous graphs, the NASA data appears to be fairly evenly distributed throughout the plot, indicating that the NASA/SEL projects, and projects recorded in the RADC database had basically similar development schedules and development time constraints for projects of similar size. As expected, each of the graphs indicate that the duration of project development correlates reasonably well with the size of a project.

Figures 4A, 4B Total Errors vs. Size

Figure 4B represents the relationship between the size of the project and the number of Software Problem Reports (SPRs) for RADC projects and Change Reports (CRs) for NASA/SEL projects; these figures representing a measure of the number of errors encountered during project development. Figure 4B represents the same relationship for HOL projects only. In this case, the addition of the NASA data resulted in a regression equation with a higher correlation coefficient than the original correlation coefficient as computed by Richard Nelson (the original being 0.583 and the most recent being 0.706). This would imply that within the RADC and NASA/SEL data, the Number of Errors is more strongly correlated to the Project Size, than shown by the projects in the RADC database alone. The two graphs here differ very little, and the standard error of the regression estimates does not improve by excluding Assembly language projects. However, both graphs indicate that for NASA and RADC projects of comparable size, the NASA projects generally have fewer errors. As expected, the number of errors encountered in a software project is fairly highly correlated to the size of the project.

Figures 5A, 5B Spatial Error Rate vs. Size

Figure 5A represents the relationship between the size of a project and the average number of errors reported during development for every 1000 lines of code, for all RADC and NASA/SEL projects, for which data was collected. The addition of the NASA data changes the regression equation somewhat, specifically reducing the resulting regression estimation. However, the standard error magnitude and degree of correlation for the regression, indicate that the Spatial Error Rate is not highly correlated to Project Size. The exclusion of projects using Assembly language code from the data set in Figure 5B reduces the standard error of the regression only slightly. Since there are only ten projects within the RADC Software Productivity Database which contain error data, and which can be considered to have been developed without the use of MPPs, any results based on these could not be considered very reliable and, consequently, an evaluation of the effect of the use of MPPs on this error rate was not examined. Although the size of the data sample is relatively small,

the NASA data points indicate a tendency for a lower error rate in those projects as compared to the projects described in the RADC database, and this is evident by the relative location of these points within the scatter plot.

Figures 6A, 6B Effort Based Error Rate vs. Size

Figure 6A represents the relationship between the size of the project and the average number of errors resulting from every ten manmonths of development effort spent on the project. This regression result is similar to the one reported in the Nelson report. This is reinforced by the observation that the NASA data is fairly evenly distributed throughout the plot. The negative coefficient of correlation indicates that, to a small degree, the size of the project correlates to a decrease in the number of errors encountered for a given amount of effort. The elimination of projects written in Assembly language in Figure 6B reduces the standard error of the regression estimate only slightly. Also, it is worth noting again, that the size of the data sample is rather small.

Figure 7A, 7B Temporal Error Rate

Figures 7A and 7B exhibit the relationship between the size of a project and the frequency or rate at which errors occur during development. This relationship was not examined in the original report by Richard Nelson. Again, the removal of projects written in Assembly language code from the data set reduces the standard error of the regression estimate only slightly. Both figures indicate that the NASA projects tended to have a lower rate of error occurence than the projects in the RADC database. The rate at which errors occur is only moderately correlated with Project Size. Again, the sample size is rather small to use as a basis to derive significant results.

3. CONCLUSIONS

Overall, it appears that RADC and NASA/SEL data are not similar. In most instances, the NASA/SEL data appeared to have a higher productivity, lower project development effort and lower error rates than the RADC productivity data. In addition, Productivity and the Number of Errors per 1000 lines appeared to be independent of total Project Size, while Development Effort, Project Duration, total Number of Errors, number of errors per unit of effort and number of errors per unit time were, to a degree, sensitive to total Project Size.

Therefore, a more indepth study of these two datasets is recommended in order to establish, statistically, the degree of this dissimilarity and to analyze its possible causes.

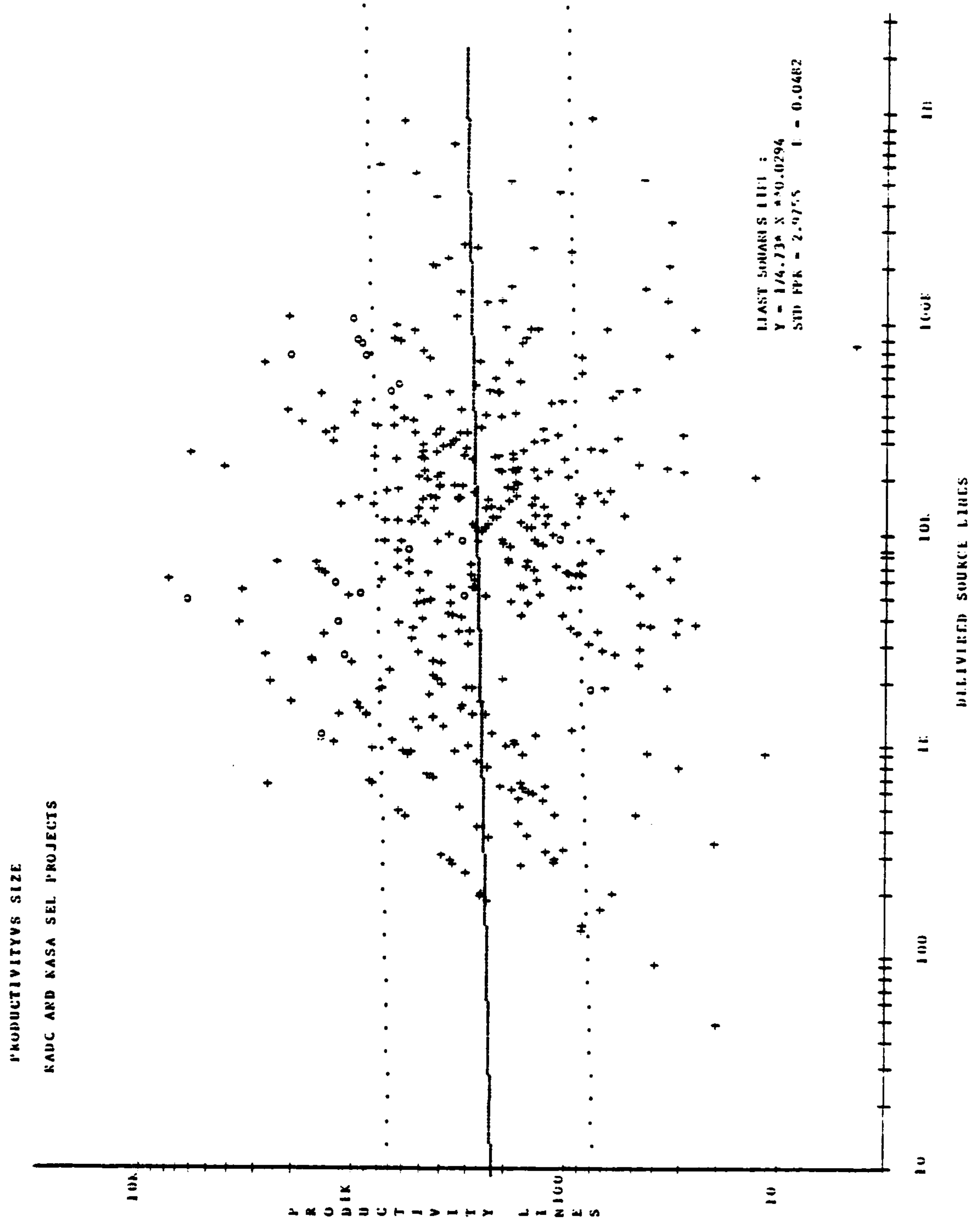

PRODUCTIVITY VS SIZE
RADC AND NASA SEL PROJECTS
DELIVERED SOURCE LINES

FIGURE 1A

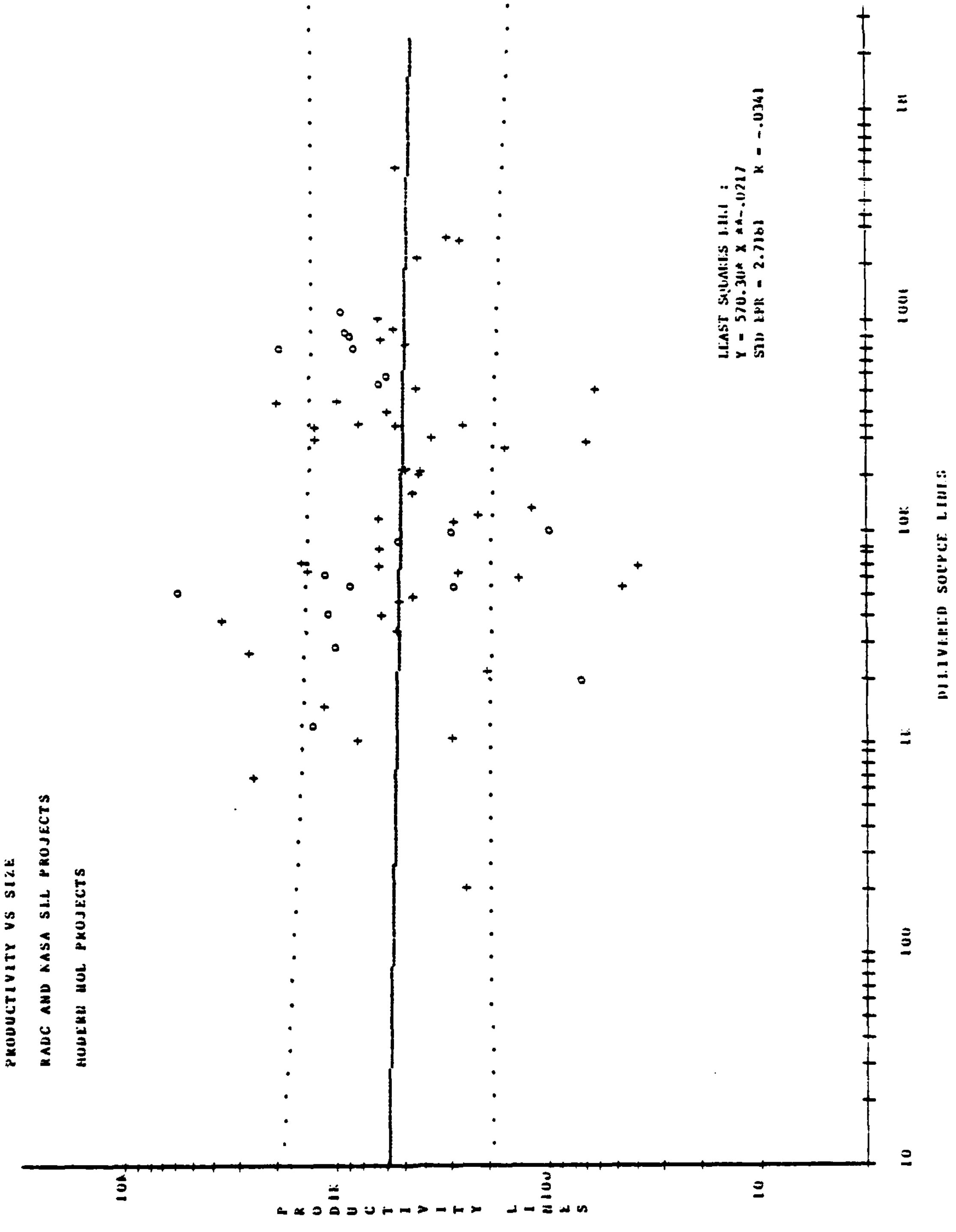

PRODUCTIVITY VS SIZE
RADC AND NASA SEL PROJECTS
MODERN MOL PROJECTS
LEAST SQUARES LINE :
DELIVERED SOURCE LINES

FIGURE 1B

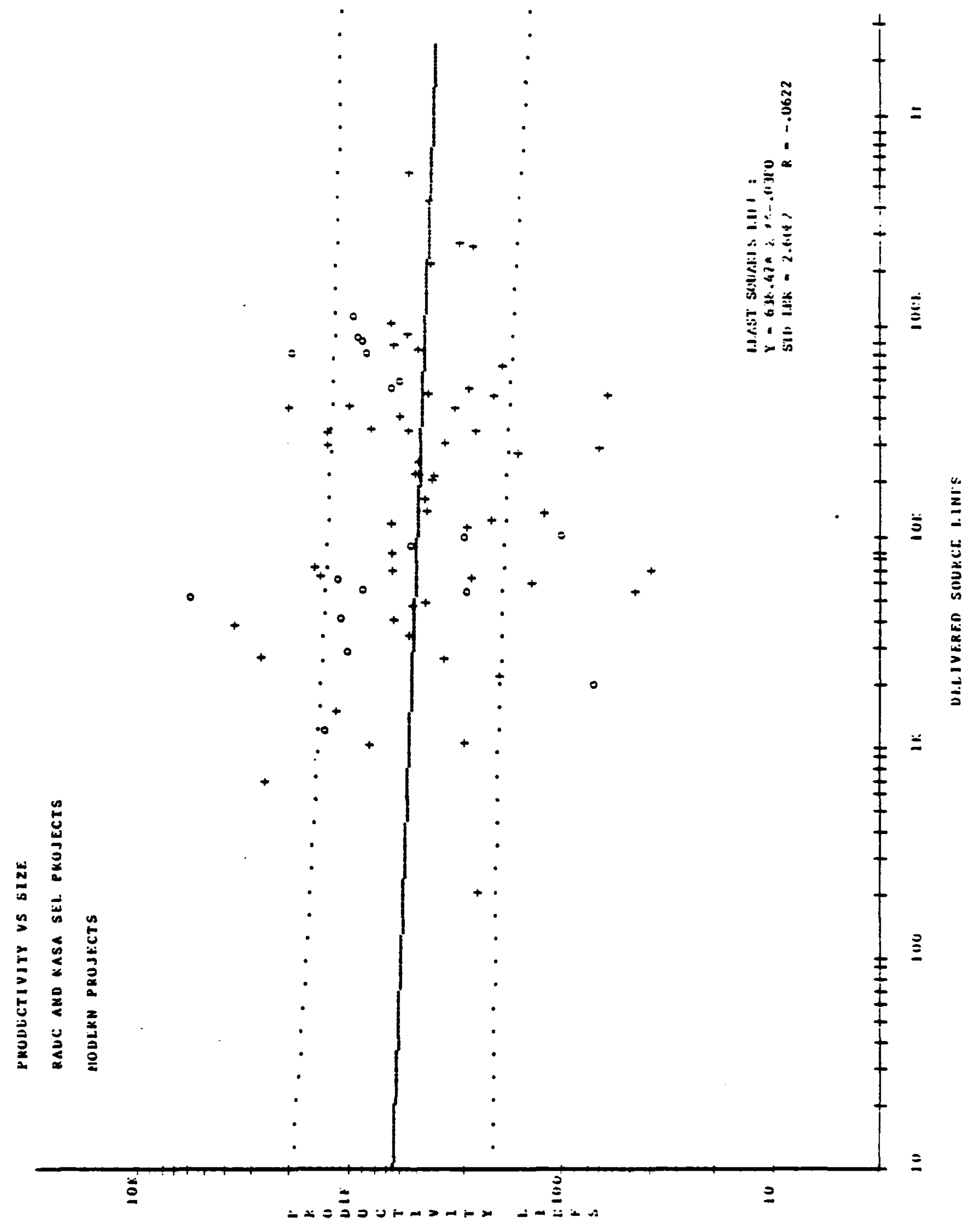

PRODUCTIVITY VS SIZE
RADC AND NASA SEL PROJECTS
MODERN PROJECTS
DELIVERED SOURCE LINES

FIGURE 1C

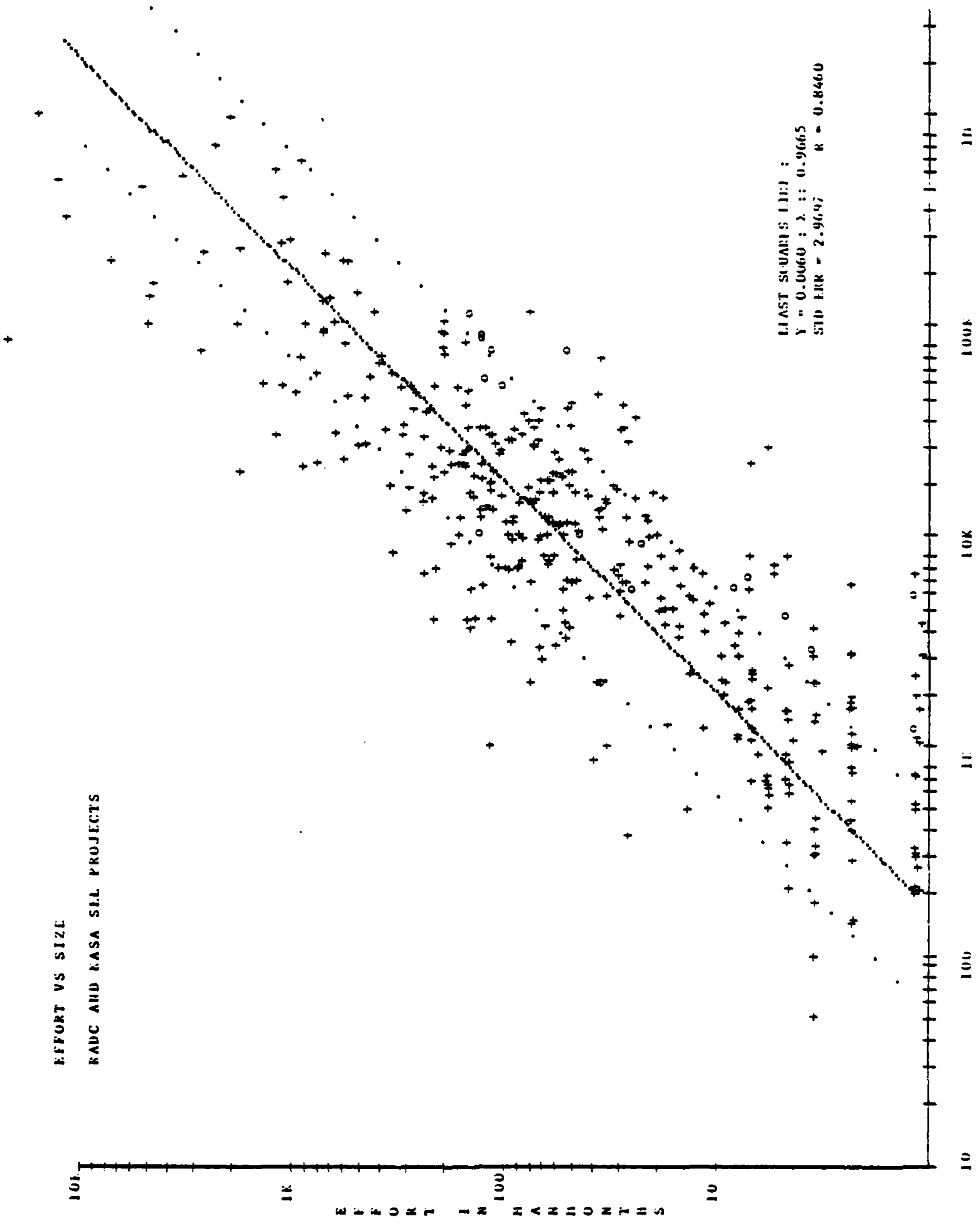

EFFORT VS SIZE
RADC AND NASA SEL PROJECTS
LEAST SQUARES LINE :
STD ERR = 2.9697
R = 0.8460
DELIVERED SOURCE LINES
EFFORT IN MANMONTHS
10
100
1K
10K
100K

FIGURE 2A

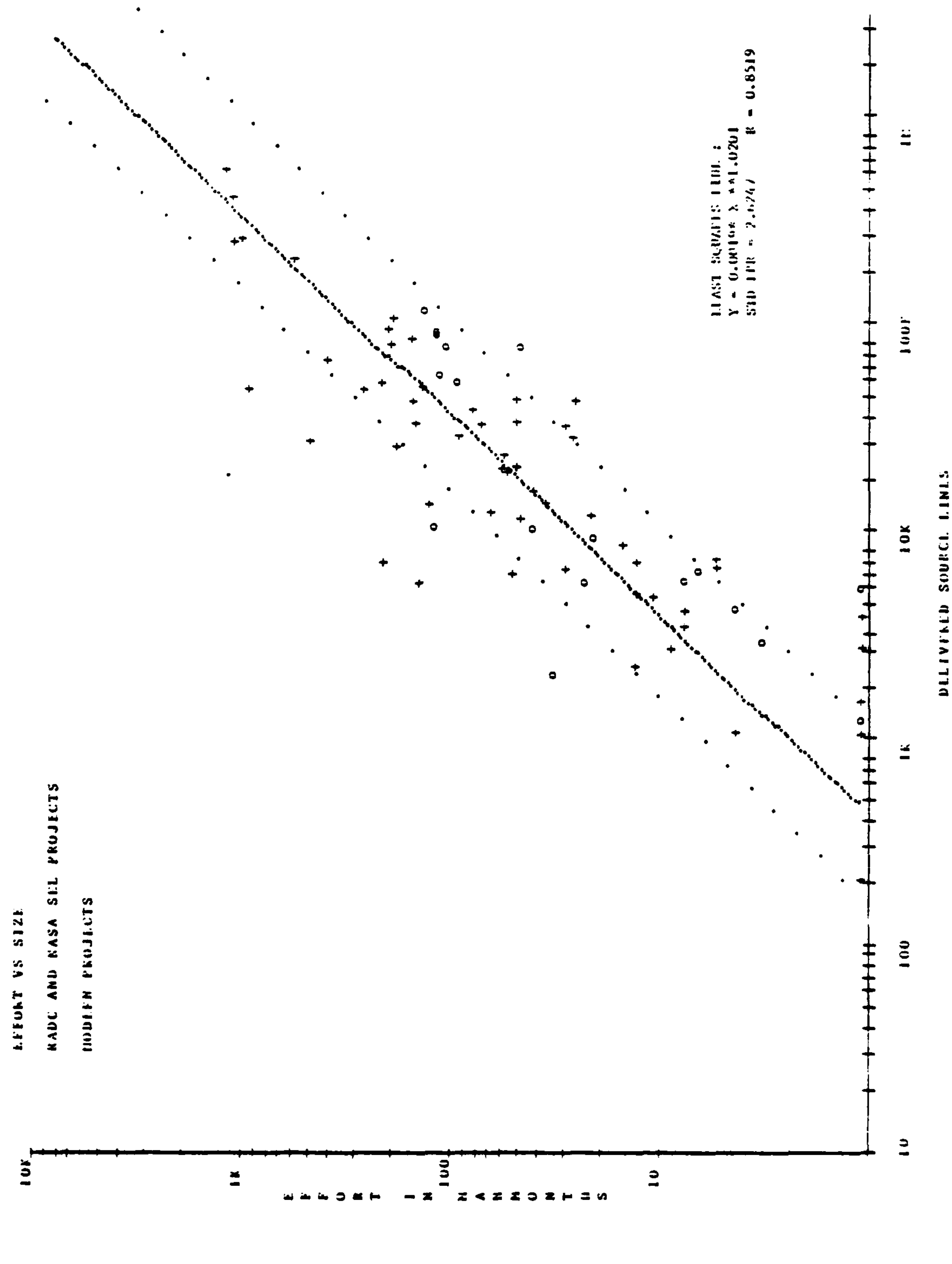

EFFORT VS SIZE
MODERN PROJECTS
DELIVERED SOURCE LINES
EFFORT IN MANMONTHS
R = 0.8519

FIGURE 2B

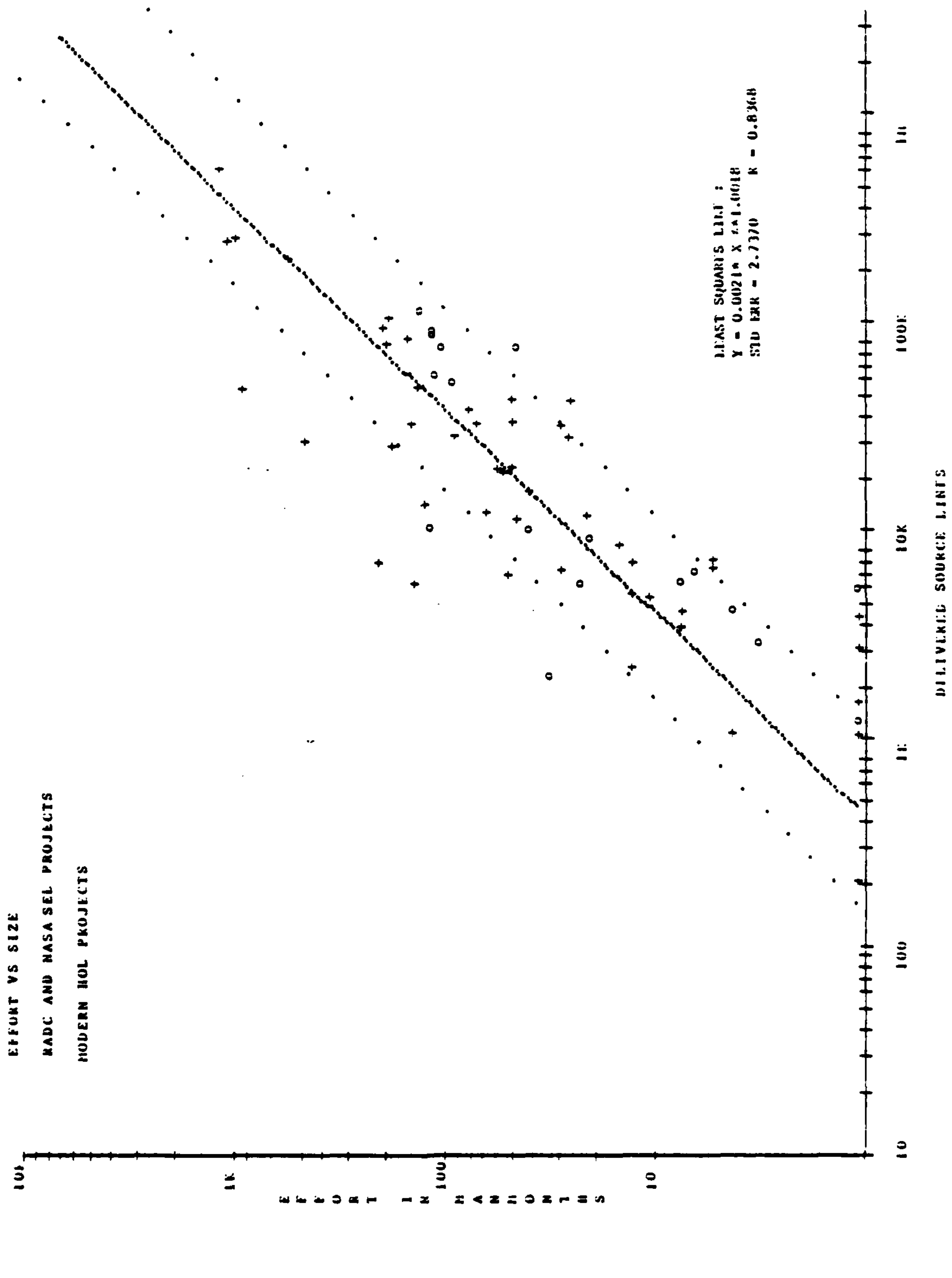

EFFORT VS SIZE
RADC AND NASA SEL PROJECTS
MODERN HOL PROJECTS
LEAST SQUARES LINE :
DELIVERED SOURCE LINES
EFFORT IN MANMONTHS

FIGURE 2C

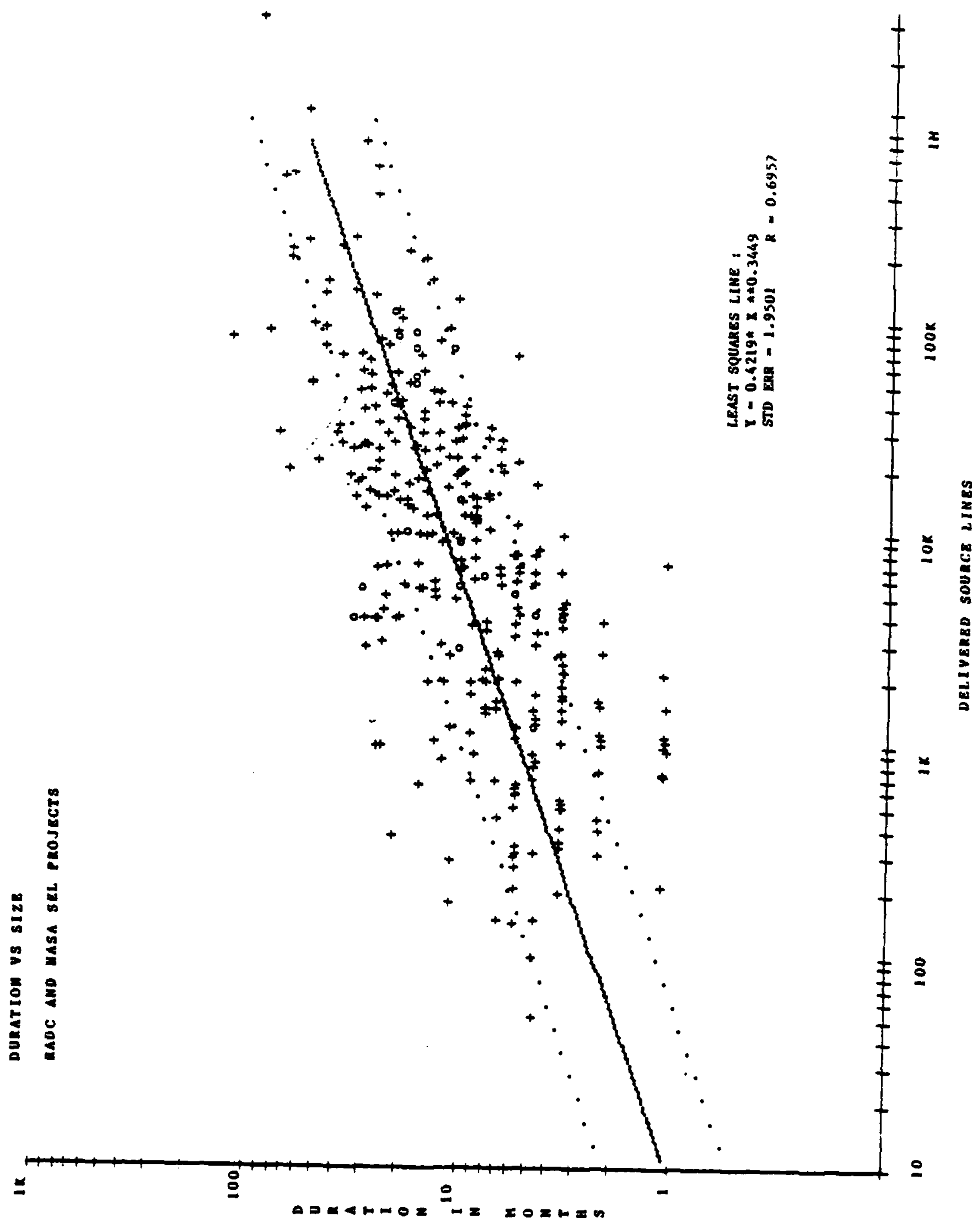

DURATION VS SIZE
RADC AND NASA SEL PROJECTS
LEAST SQUARES LINE :
Y = 0.4219* X **0.3449
STD ERR = 1.9501
R = 0.6957
1K
100
10
1
DURATION IN MONTHS
10
100
1K
10K
100K
1M
DELIVERED SOURCE LINES

FIGURE 3A

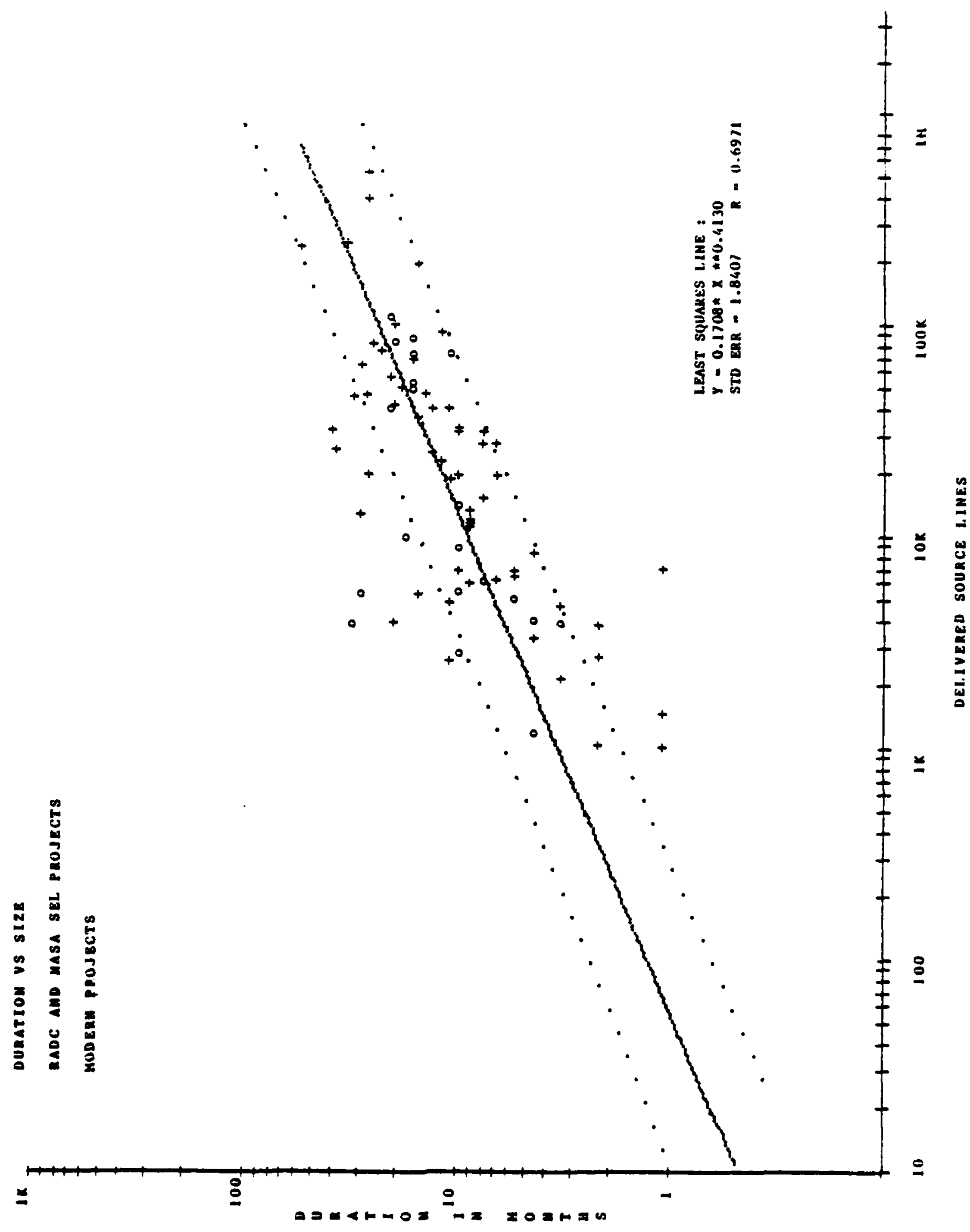

DURATION VS SIZE
RADC AND NASA SEL PROJECTS
MODERN PROJECTS
LEAST SQUARES LINE :
Y = 0.1708* X **0.4130
STD ERR = 1.8407 R = 0.6971
DURATION IN MONTHS
1K
100
10
1
DELIVERED SOURCE LINES
10
100
1K
10K
100K
1M

FIGURE 3B

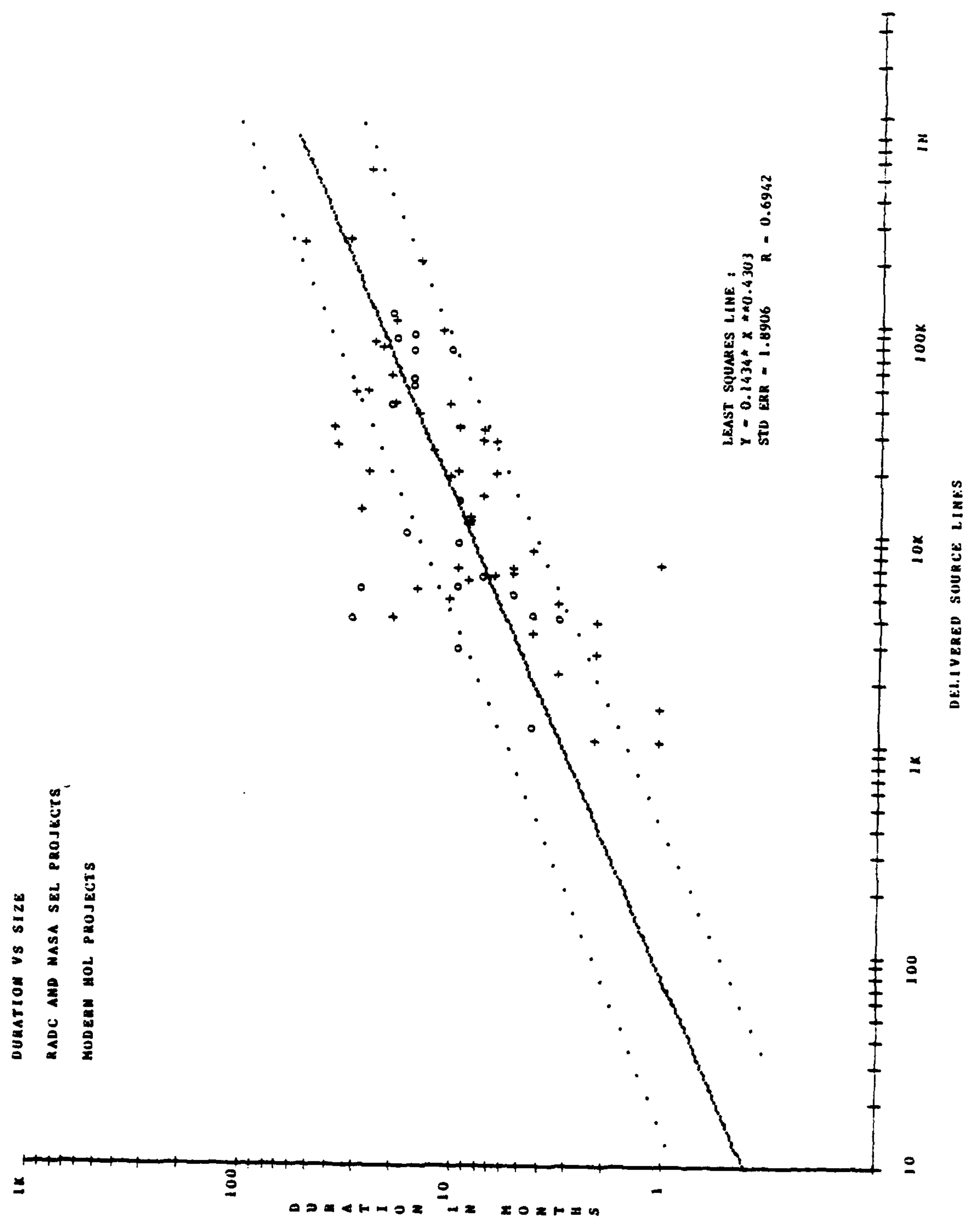
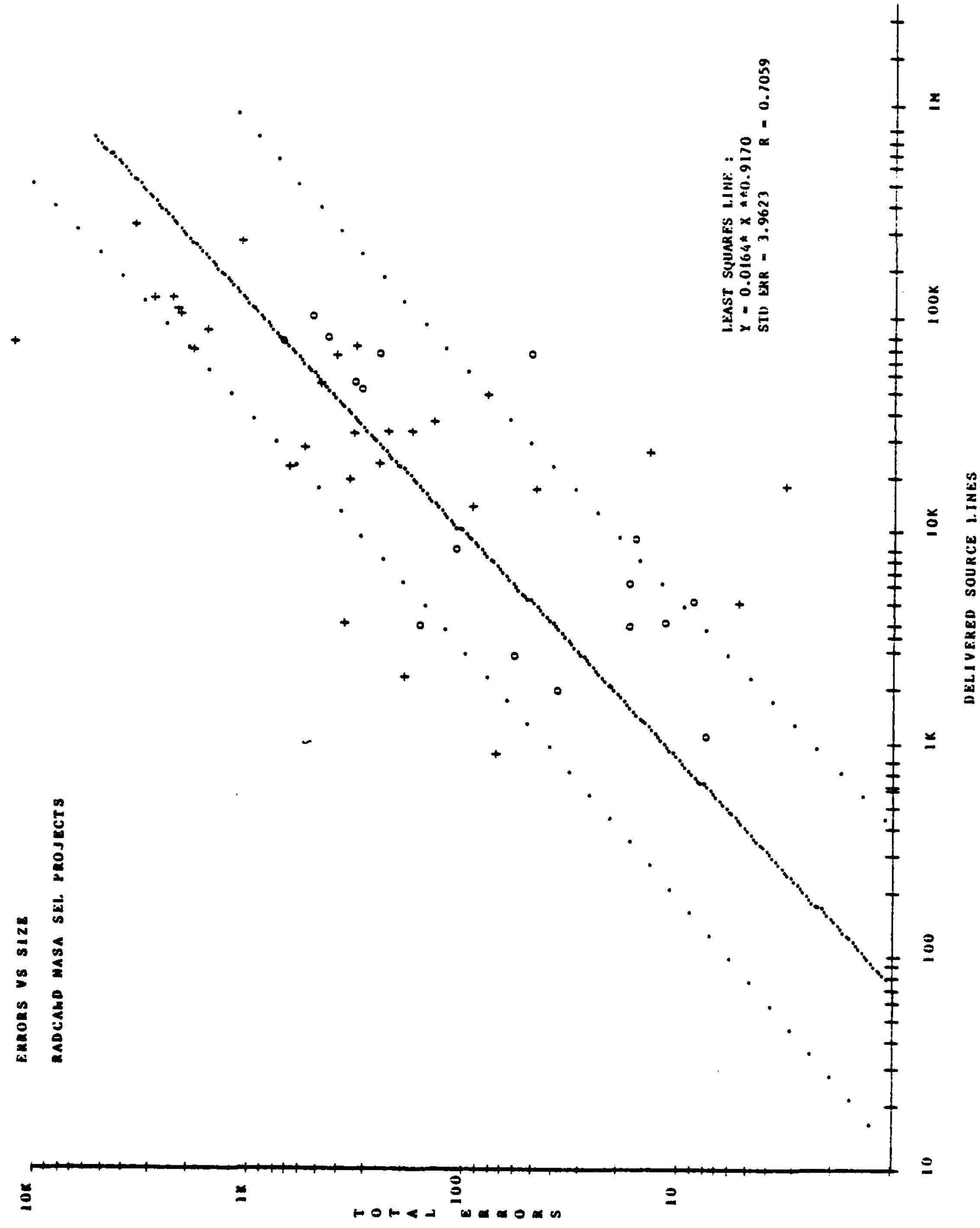

DURATION VS SIZE
RADC AND NASA SEL PROJECTS
MODERN HOL PROJECTS
LEAST SQUARES LINE :
Y = 0.1434* X **0.4303
STD ERR = 1.8906 R = 0.6942
DURATION IN MONTHS
1K
100
10
1
DELIVERED SOURCE LINES
10
100
1K
10K
100K
1M

FIGURE 3C



ERRORS VS SIZE
RADCAND NASA SEL PROJECTS
LEAST SQUARES LINE :
Y = 0.0164* X **0.9170
STD ERR = 3.9623 R = 0.7059
10K
1K
100
10
TOTAL ERRORS
10
100
1K
10K
100K
1M
DELIVERED SOURCE LINES

FIGURE 4A

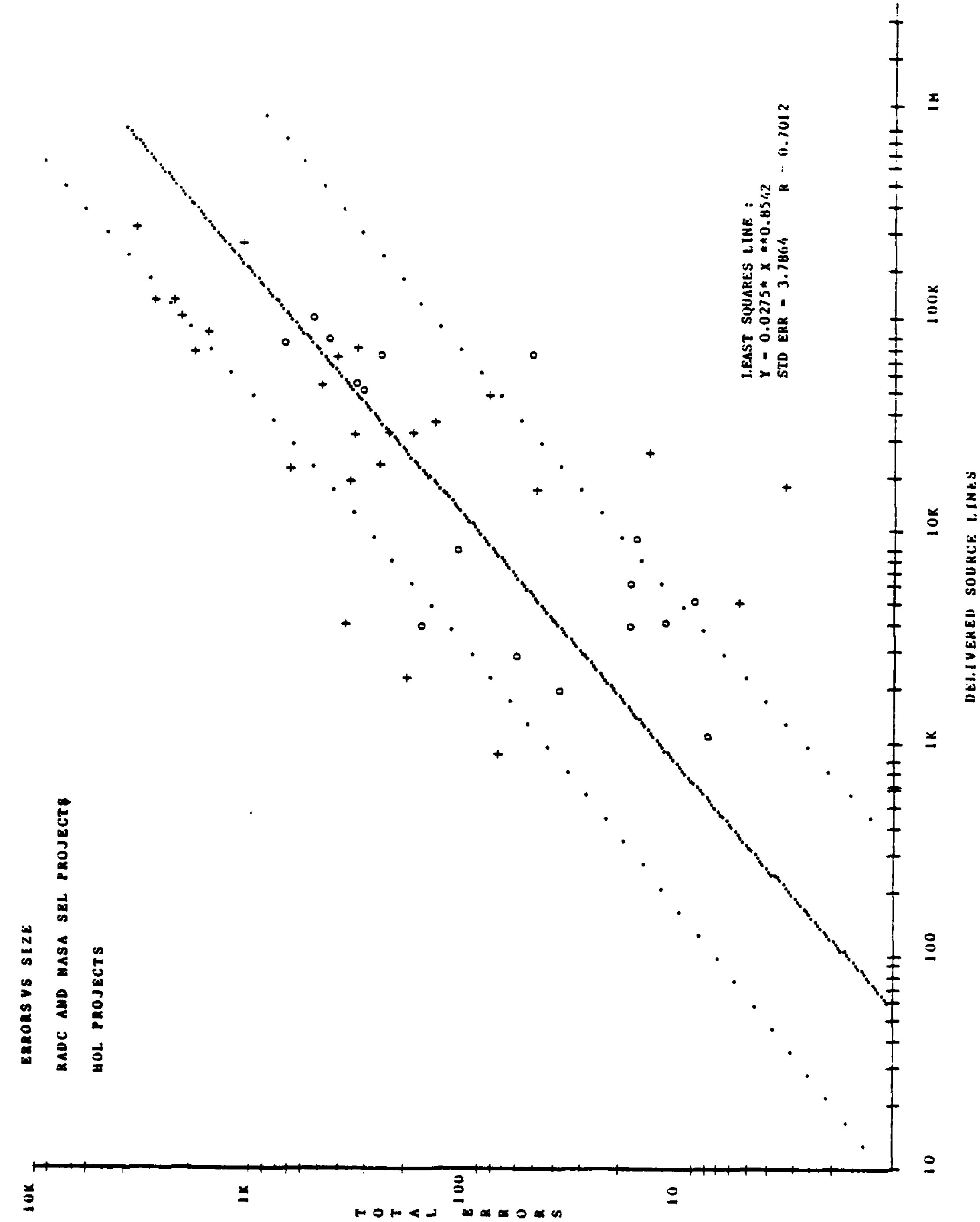

ERRORS VS SIZE
RADC AND NASA SEL PROJECTS
HOL PROJECTS
LEAST SQUARES LINE :
Y = 0.0275* X **0.8542
STD ERR = 3.7864 R = 0.7012
10K
1K
100
10
TOTAL ERRORS
10
100
1K
10K
100K
1M
DELIVERED SOURCE LINES

FIGURE 4B

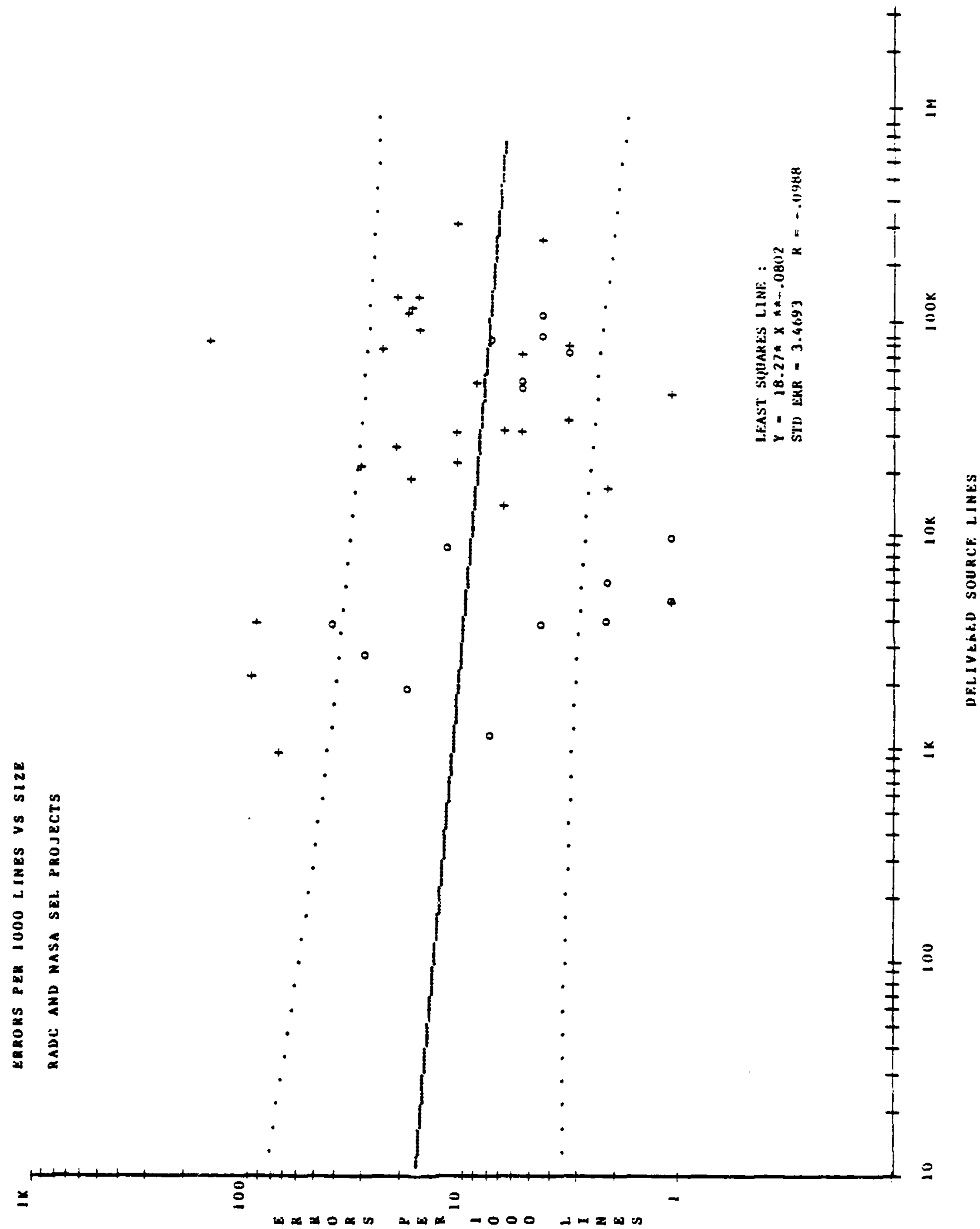

ERRORS PER 1000 LINES VS SIZE
RADC AND NASA SEL PROJECTS
LEAST SQUARES LINE :
Y = 18.27* X **-.0802
STD ERR = 3.4693 R = -.0988
DELIVERED SOURCE LINES
ERRORS PER 1000 LINES
10
100
1K
10K
100K
1M
1K
100
10
1

FIGURE 5A

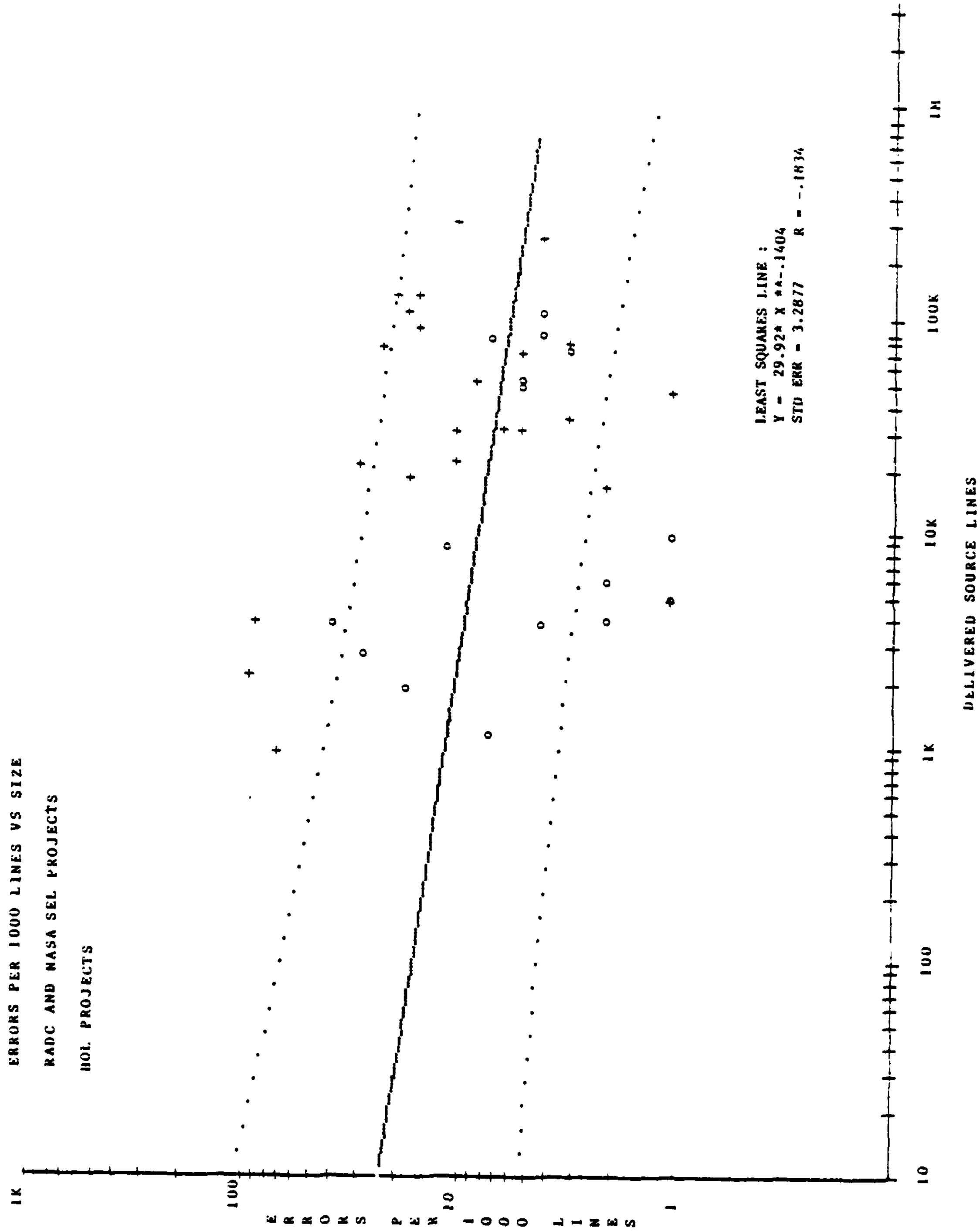

ERRORS PER 1000 LINES VS SIZE
RADC AND NASA SEL PROJECTS
HOL PROJECTS
LEAST SQUARES LINE :
Y = 29.92* X **-.1404
STD ERR = 3.2877 R = -.1834
DELIVERED SOURCE LINES
ERRORS PER 1000 LINES
1K
100
10
1
10
100
1K
10K
100K
1M

FIGURE 5B

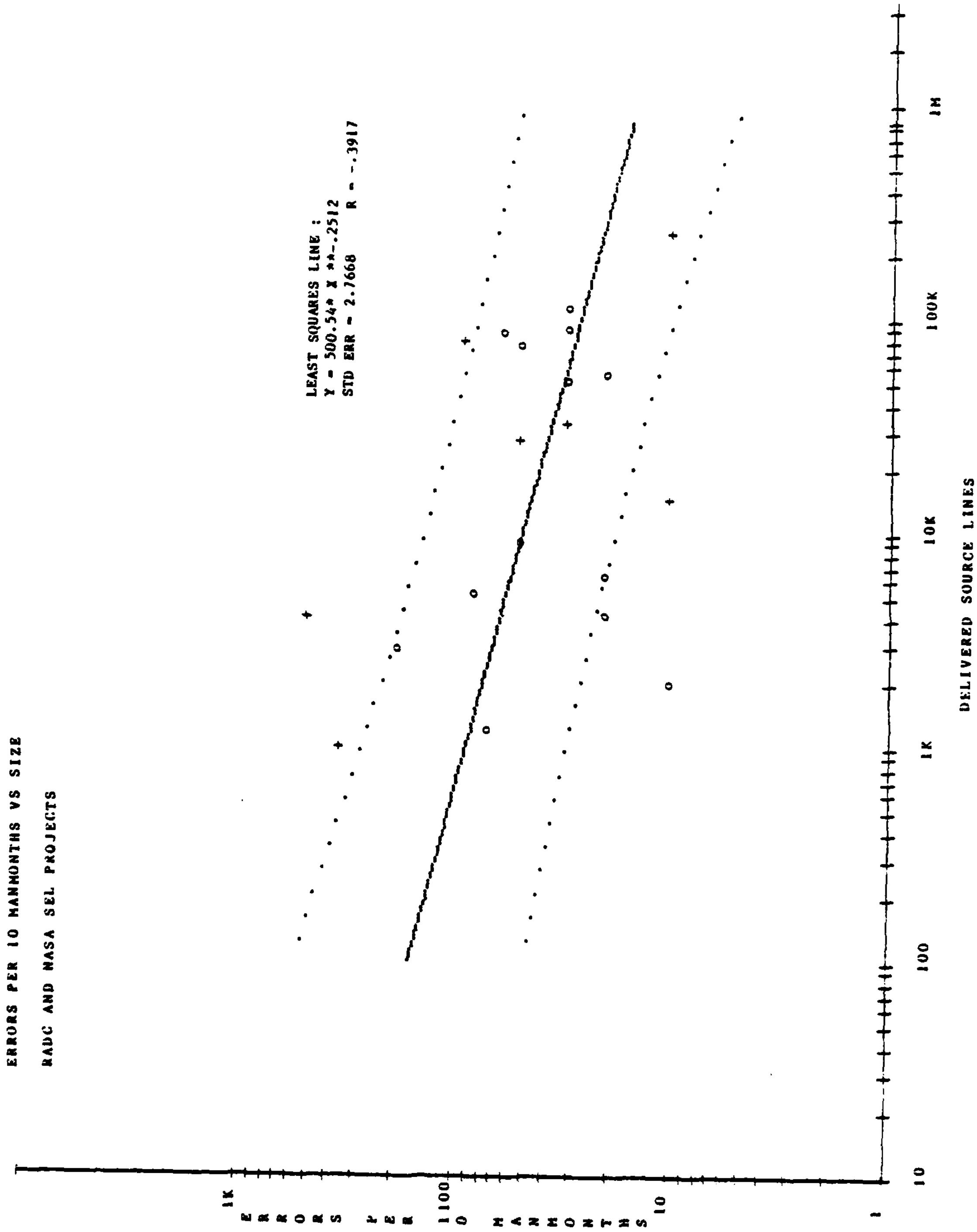

ERRORS PER 10 MANMONTHS VS SIZE
RADC AND NASA SEL PROJECTS
LEAST SQUARES LINE :
Y = 500.54* X **-.2512
STD ERR = 2.7668 R = -.3917
ERRORS PER 10 MANMONTHS
1K
100
10
1
10
100
1K
10K
100K
1M
DELIVERED SOURCE LINES

FIGURE 6A

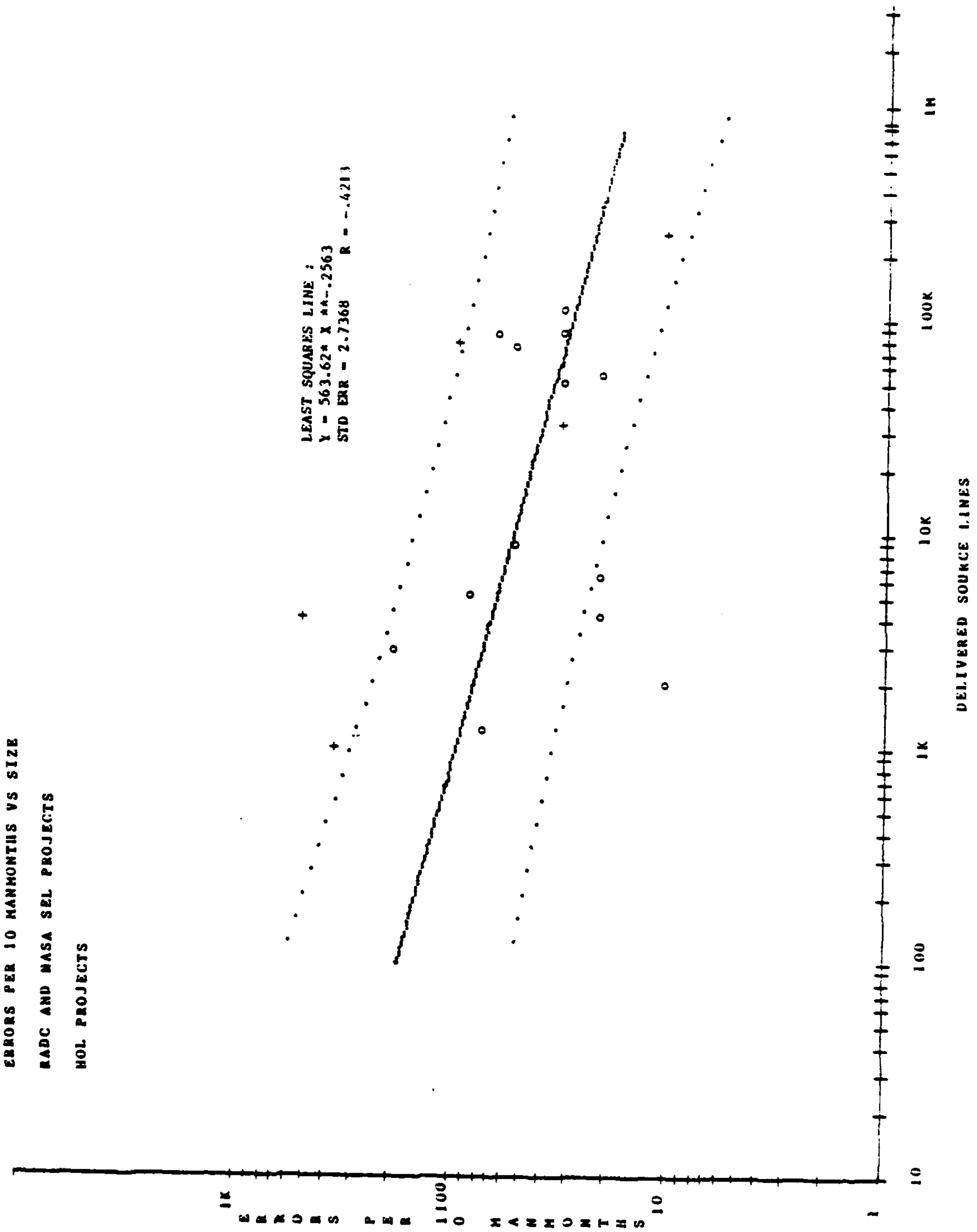

ERRORS PER 10 MANMONTHS VS SIZE
RADC AND NASA SEL PROJECTS
HOL PROJECTS
LEAST SQUARES LINE :
Y = 563.62* X **-.2563
STD ERR = 2.7368
DELIVERED SOURCE LINES
ERRORS PER 100 MANMONTHS
10
100
1K
10K
100K
1M
1

FIGURE 6B

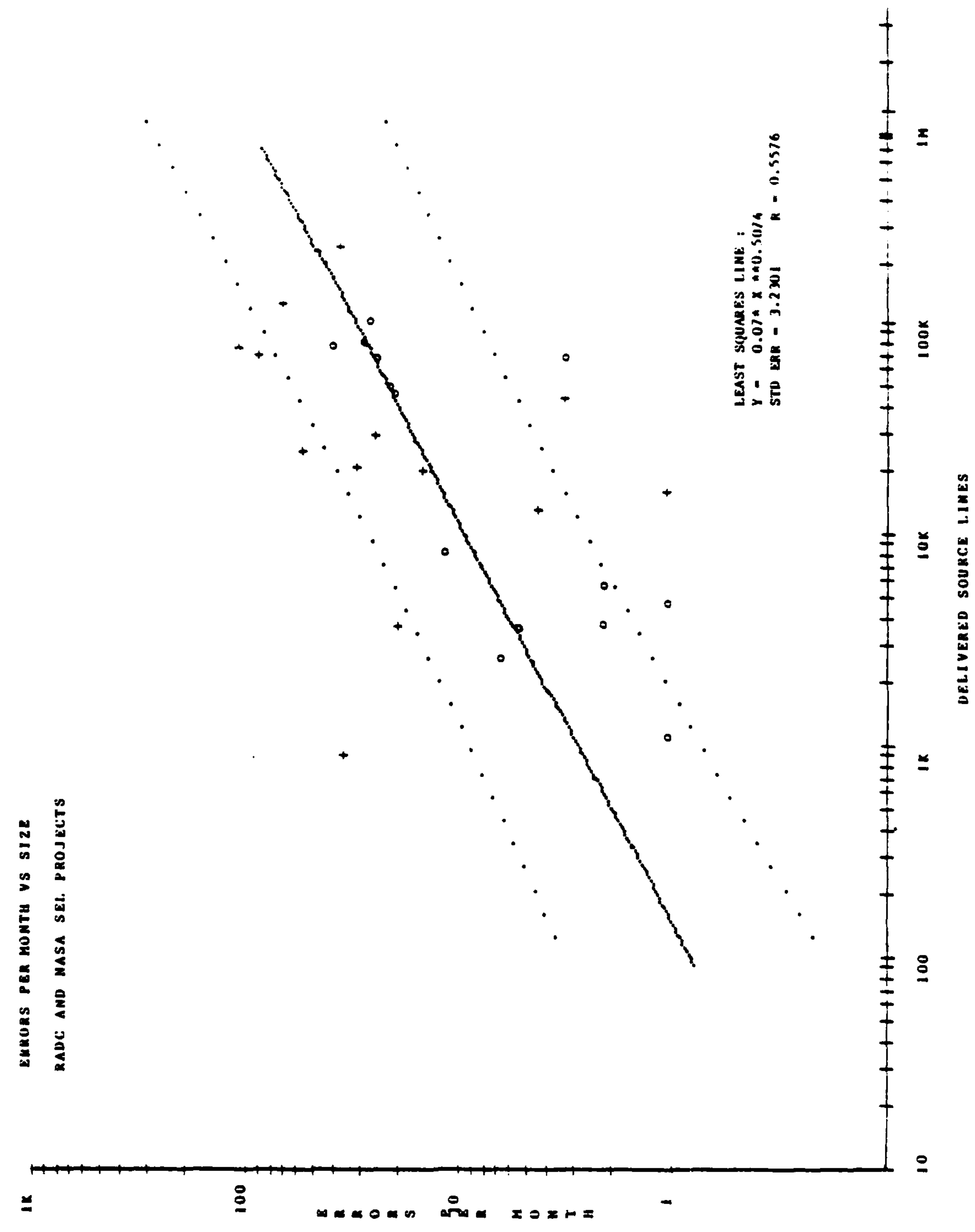

ERRORS PER MONTH VS SIZE
RADC AND NASA SEL PROJECTS
LEAST SQUARES LINE :
Y = 0.074 X **0.5074
STD ERR = 3.2301 R = 0.5576
ERRORS PER MONTH
1K
100
10
1
10
100
1K
10K
100K
1M
DELIVERED SOURCE LINES

FIGURE 7A

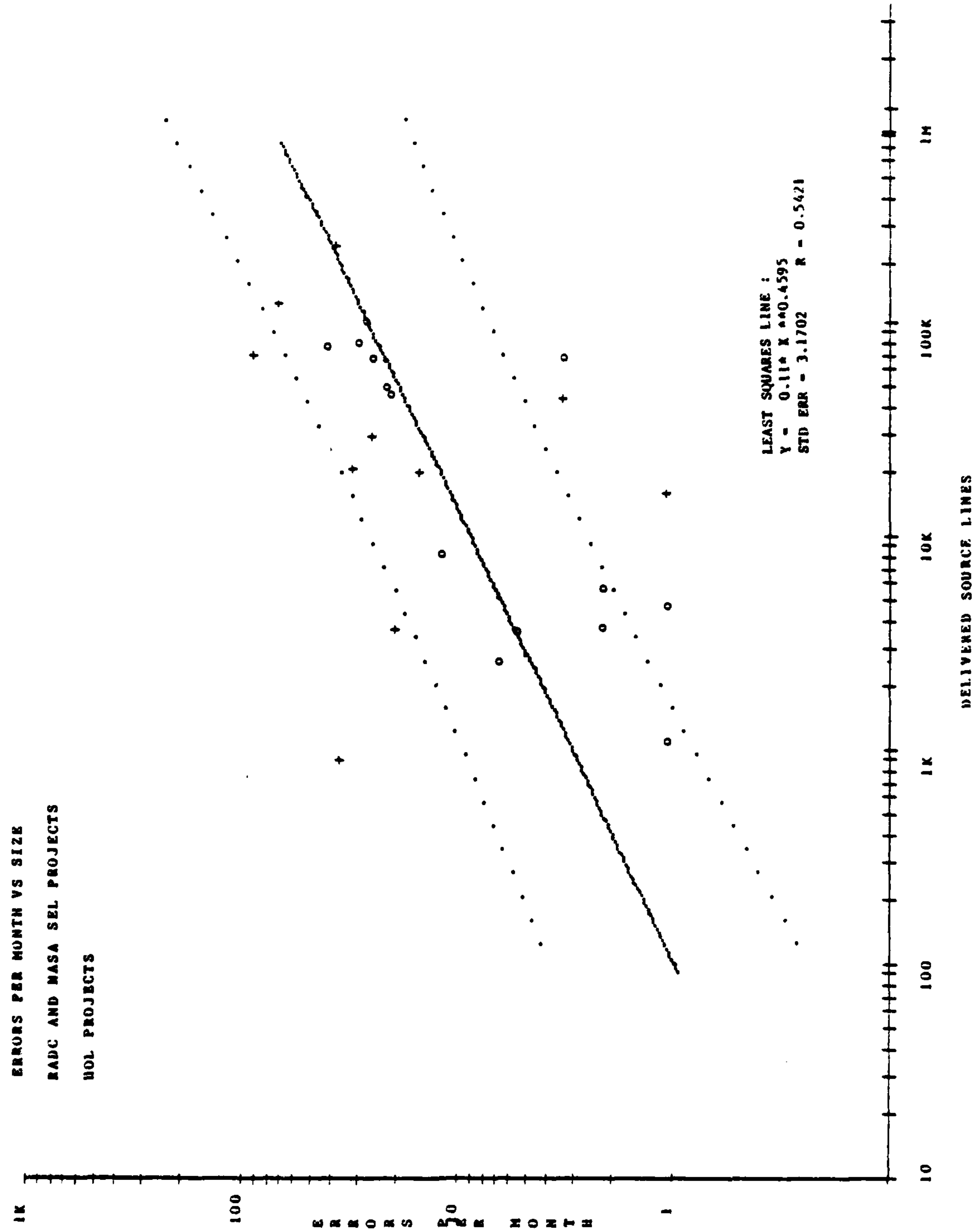

ERRORS PER MONTH VS SIZE
RADC AND NASA SEL PROJECTS
HOL PROJECTS
LEAST SQUARES LINE :
Y = 0.11* X **0.4595
STD ERR = 3.1702 R = 0.5421
DELIVERED SOURCE LINES
ERRORS PER MONTH
10
100
1K
10K
100K
1M
1K
100
10
1

FIGURE 7B

APPENDIX A

FIGURE A1

PRODUCTIVITY VS DSLOC
(CONVENTIONAL PROGRAMS)

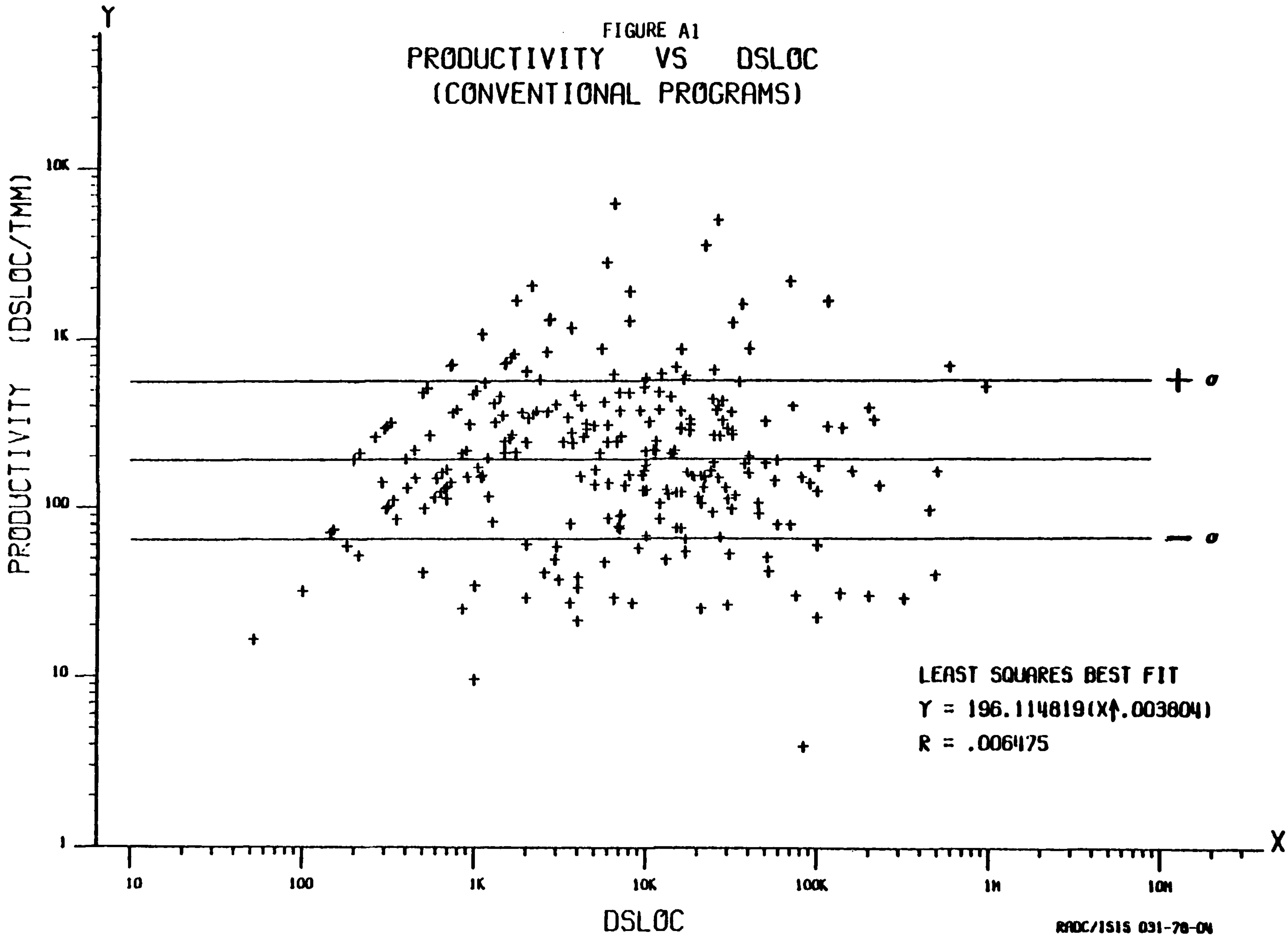


FIGURE A2

TOTAL MAN MONTHS VS DSLOC
(CONVENTIONAL PROGRAMS)

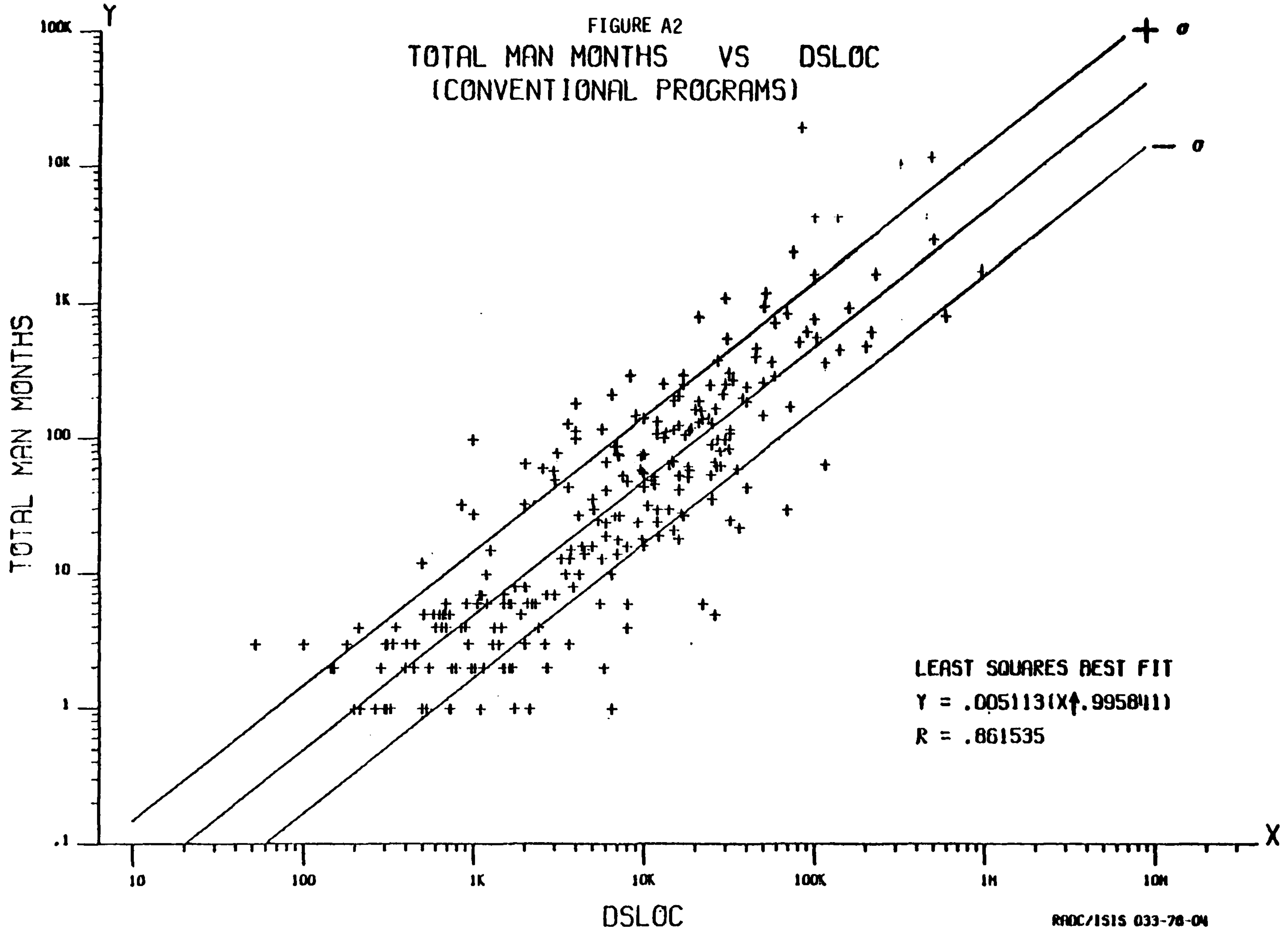


FIGURE A3

PROJECT DURATION (MONTHS) VS DSLOC
(CONVENTIONAL PROGRAMS)